Project	AtlantOS – 633211
Deliverable number	D7.1
Deliverable title	Data Harmonization Report: Report containing recommendation on data harmonization
Description	Report harmonization in data and data processing to facilitate the interoperability of the systems.
Work Package number	7
Work Package title	Data flow and data integration
Lead beneficiary	UniHB
Lead authors	Koop-Jakobsen (UniHB), Waldmann (UniHB), Huber (UniHB), Harscoat (Ifremer), Pouliquen (Ifremer)
Contributors	All WP7 networks and integrators
Submission data	N/A
Due date	30.09.2016
Comments	N/A

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement nº 633211.

Last updated: 04 October 2016

Table of Contents

Executive summary:

AtlantOS WP7 is dedicated to improve harmonization of data management procedures, and thereby improve the quality, interoperability and discoverability of data resources in AtlantOS. To improve harmonization, AtlantOS WP7 works on multiple levels;

a) WP7 has identified selected areas, where significant improvements of interoperability can be obtained. This has resulted in the formulation of a common agreement stating a set of specific minimum standards, which shall ensure cross platform coherence. This includes minimum standards for use of identifiers for platforms and institutions, metadata including vocabularies, quality control and dissemination means. Furthermore, guidelines regarding DOI assignment, catalogue techniques and vocabulary use in AtlantOS have been formulated.

b) AtlantOS has formulated and installed a Data Management Plan (DMP) setting the framework for handling and dissemination of AtlantOS data. This was the first step towards improved harmonization and includes an overview of the Data Landscape, prioritization of Essential Variables for AtlantOS, regulations regarding open access to data and recommendations on use of standards.

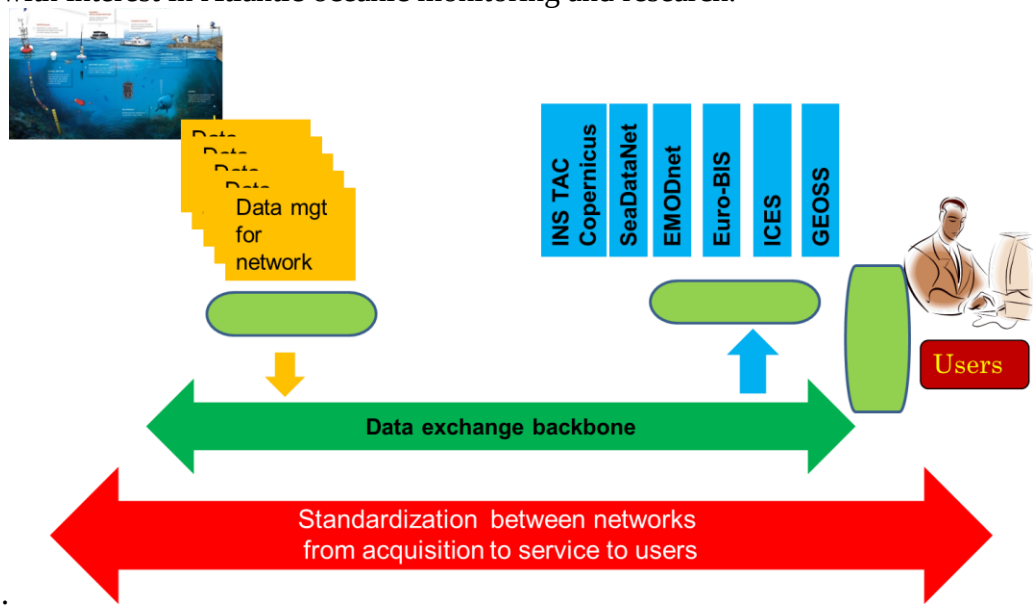
c) AtlantOS WP7 is initiating investigations of the use of GEOSS services, both for technical broker solutions to improve harmonization as well as for dissemination of AtlantOS data resources in an interdisciplinary global context.

d) AtlantOS is also working on improving the transcontinental data sharing. A workshop is planned for in 2017 specifically targeting improvement of transcontinental sharing of data from the Atlantic Ocean. We here present the preliminary incentives for improving the transatlantic collaboration.

Introduction:

AtlantOS WP7 involves data-providing in-situ observing **Network** systems (GEOMAR for GOSHIP and Sea floor Mapping, University of Exeter for SOOP, SAHFOS for CPR, ICES for Fish+plankton, Ifremer for Argo, NERC for OceanSites, CNRS for Glider, EUMETNET for Surface drifter, HZG for Ferry box, BODC for Tide gauges), as well as data infrastructures integrating in-situ observations from existing European and international data providers (INSTAC Copernicus, SeaDataNet, EMODnet, ICES, EurOBIS, GEOSS) so called **Integrators**.

The work of AtlantOS WP7 is dedicated to improve the data management and interoperability among the observation networks and Integrators involved in AtlanOS. WP7 shall ensure availability of the diverse and interdisciplinary data pool collected from the Atlantic Ocean by different ocean observing platforms in AtlantOS. The overall goal is to make these data freely available, in a readily usable format, with sufficient information attached in the form of metadata for a scientific use as well as monitoring purposes. These data resources shall be publically available to all stakeholders with interest in Atlantic oceanic monitoring and research.



.

Figure 1: The Data system for AtlantOS. The AtlantOS networks shall feed their information to the Data integrators through a Data exchange backbone. AtlantOS WP7 is in task 1+2 setting up the framework for the data exchange backbone.

The data networks and data integrators in the AtlantOS project are overall mature networks with long-term experience in the collection, handling, curation and dissemination of data and meta-data. In this regard, the data-networks overall have well-established work-flows and policies for their data-management. Consequently, trying to implement a sovereign and rigid set of regulations for standards and workflows for all the networks and integrators in AtlantOS to comply with, would be highly challenging and not in the best interest of AtlantOS. The scope of WP7 is therefore to explore the data landscape and hereby identify needs for improvements in regard to harmonization and standardization, which will be to the benefit of all data-networks and their connection to the data integrators involved in WP7. In this regards, significant progress has been made on the following selected standardization tasks:

1) Identifying the data landscape and prioritizing a list of EOVs across the Networks involved in WP7.

2) Identifying a minimum set of metadata common vocabularies to be used by all networks.
3) Giving recommendations for providing a minimum level of Near Real Time Quality Control Procedures for selected EOVs (T, S, Current, O_2, Chl, Nitrate, Sea Level, Carbon).
4) Identifying and improving gabs or impediments in basic services for discovery, viewing and downloading of data.

This report describes the initiatives generated under WP7-task 1 to ensure improved harmonization of data resources in AtlantOS. Furthermore, this report shall look at the data management plan formulated under AtlantOS as well as recommend guidelines that shall improve the potential for harmonization of data resources across the Atlantic.

WP7 initiatives for harmonization among AtlantOS data networks

AtlantOS Data landscape and EOVs

The data networks in AtlantOS offers a wide variety of marine data related to many different scientific disciplines. The data ranges from standard parameters such as common physical ocean measures, such as conductivity, temperature and density to specialized variables such as isotopes of O_2, N_2, and Fish and Plankton surveys (ICES). This results in a vast pool of very heterogeneous data collected on different spatial and temporal scales and with different instrumentation. The overall data landscape in AtlantOS was incorporated into the AtlantOS data management plan (page 3-7) (See Appendix 1) (AtlantOS_DMP 2015).
The recent decade's rapid development of marine technology allows for the deployment of more and more autonomously operated observation systems. This opportunity to collect almost unlimited amounts of data has also accommodated a significant need for a prioritization of parameters measured in the global oceans as well as in the Atlantic. With the purpose of optimizing the use for funding resources for automated ocean monitoring, various groups and organizations have in recent years debated the identification of Essential Ocean Variables (EOV) for physics, biogeochemistry and biology, ecosystems as part of the Framework on Ocean Observing (GOOS).
Essential Variables for AtlantOS was identified as Temperature, Salinity, Currents, Oxygen, Chlorophyll, Nitrate, Sea Level and Carbon. The prioritization of variables were incorporated into the AtlantOS data management plan (AtlantOS_DMP 2015) available at https://www.atlantos-h2020.eu/download/deliverables/11.2%20Data%20Management%20Plan.pdf and a synthesis is presented in Appendix 1.

Metadata and Vocabularies

Networks and Integrators agreed on a minimum and essential set of applicable recommendations relying on existing international standards that will ensure cross platform coherence and also facilitate data discovery for users and data integration:

- **Platforms** should have a unique identifier that will be either WMO for most platforms or ICES code for ships
- **Metadata** used by the networks for parameters should be "mappable" on standard vocabularies existing and EU (SeaDataNet vocabularies) or international (CF or WoRMS for Taxa). A vocabulary matrix for AtlantOS EOVs was built and validated with the Network representatives, it is described in appendix 2 and accessible here: https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search/A05/

- **Institutions** used in a data file shall be identified by a unique code from the EDMO (European Directory of Marine Organizations) existing catalogue. EDMO shall be enhanced with the Networks needs.
- **QC information** will be attached to the data; both Quality flags that can be mapped to SeaDataNet flag scale (available in the SeaDataNet Common Vocabularies (http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx/) as list L201), and, whenever known, processing level information ("qualified in RT using automated procedures" or "processed in delayed DM by Scientist")
- **Distribution means**: provide an FTP service at the level of network data management as the minimum delivery service. Additional services such as WEB services can also be provided, but are not mandatory.

Concerning the metadata for platform type and sensors, it was agreed that it was an issue to be solved at Network level and that harmonization across networks was not seen as a priority. Nevertheless, a recommendation to implement SensorML for sensors whenever possible will be issued in partnership with other projects such as FIXO3, ODIP2, ENVRI+, SeaDataCloud.

Near Real Time Quality Control Procedures for selected EOVs

Quality assurance and quality control is an essential part of data management. In the classical scientific sense quality control is inevitably the responsibility of the lead PI for any research project or monitoring program. However, with the increasing amount of autonomously collected data, the intellectually based human quality control is no longer a sufficient process for the large load of data collected in all areas of the ocean, and technical QC procedures must be installed. In particular, real-time or near-real time data constitute a challenge, and sufficient time must be given to allow for QC procedures to validate the data at different levels before further processing and/or dissemination.
In regard to real-time or near-real time data, flagging system algorithmically observing the data flow can be installed to raise awareness of abnormalities in the data signals. However, besides the technical quality control, QC-procedures also entail the completeness of metadata, consistence of formats, and correctness of download. For the selected EOVs in AtlantOS, best practice procedures for automatic assurance of a minimum level of Near Real Time Quality has been formulated by experts and validated by the Network representatives. These are compiled in [D7.2 QC Report] https://www.atlantos-h2020.eu/download/deliverables/7.2%20QC%20Report.pdf

Gaps or impediments in basic services for discovery, viewing and downloading

Throughout the work of AtlantOS WP7, there was a focus on improving basic services for management of AtlantOS resources, and wherever possible, gaps have been identified and sought closed through the formulation and documentation of best practises guidelines. This goes for catalogue techniques, assignment of Digital Object Identifiers (DOIs) assignment and network descriptions and is fully documented in the [D7.4 Data Management Handbook] https://www.atlantos-h2020.eu/download/deliverables/7.4%20Data%20Management%20Handbook.pdf
To summarize the recommendations/guidelines released:

- A document was formulated recommending a catalogue technique to be used at network global data assembling level to facilitate the discovery of platforms and data files. This technique is already implemented at the Copernicus in-situ TAC. The document is available here: http://dx.doi.org/10.13155/45063
- A document describing the general principles of Digital Object Identifiers (DOI) was formulated. It also provides two examples of DOI implementation (Argo and Cruises) being

guidelines for AtlantOS networks. A DOI is an allocation of unique identifier for data or data sets. Generally used to identify scientific publications, a DOI can be attributed to any physical, numerical or abstract resource. The document is available here: http://dx.doi.org/10.13155/44515 . The discussion on best practises for DOI assignment continues under AtlantOS as well as in other fora, where AtlantOS representatives participate, such as the Research Data Alliance. Each network in AtlantOS will propose a strategy for DOI implementation on the data it manages.

- As part of the data landscape analysis, it was found that a best practice guideline for the description of the data networks and products in AtlantOS would improve the overview of AtlantOS resources. Hence, a template for network descriptions was distributed and all Networks have first to provide their data collections description to populate an AtlantOS catalogue implemented with the geonetwork component of the **Sextant** Spatial Data Infrastructure. This template will be extended to describe products, including those issued from the WP7.5 task. Full template description is available in D7.4 https://www.atlantos-h2020.eu/download/deliverables/7.4%20Data%20Management%20Handbook.pdf

Data Harmonization in an interdisciplinary perspective

Ultimately the scope of large-scale intercontinental and interdisciplinary science projects like AtlantOS is to provide knowledge and information in the form of large pools of data to address multifaceted societal challenges by providing resources from multiple sub-disciplines. Combined these resources can help provide a holistic interpretations of pressing environmental challenges such as ecosystem responses to global climate change, and its threats such as loss of biodiversity. The AtlantOS project covers a heterogeneous pool of data, which can be used for this purpose. Integration of such diverse resources requires harmonization. Efforts to standardize research data in Oceanic research have been ongoing on national as well as multi-national levels for many years through various international projects, such as COOPEUS/COOP+, ENVRI /ENVRI+ and many others. Although, the interoperability among research infrastructures in the oceanic community has improved significantly, challenges still remain. New types of resources (e.g. the ever growing pool of genetic data) must be implemented, and interoperability in a larger perspective must be sought; for example with research infrastructures also handling and providing data of terrestrial origin. Hence, the effort to improved interoperability continues under various global initiatives such as the Research Data Alliance, GEOSS and ICSU-WDS, and also in the AtlantOS project.
Apart from the work improving the interoperability among data networks and data integrators with in AtlantOS as shown previously, the AtlantOS project also explores the broader perspective by looking into the advantages of novel web-based services, such as the GEOSS discovery and access broker can give, and by giving recommendations on harmonization guidelines that will assure the use of AtlantOS resources in a global perspective.

Data harmonization through brokering approaches

The rapid developments of brokering approaches by data integrators to a large extend obviate the tedious and cumbersome implementation of rigid sets of standards. Major Data integrators are working on developing brokers in order to improve cross platform interoperability and lower entry barriers for both users and data-providers. In a fully operational broker-approach, users as well as data providers will not be asked to comply with specific standard regulations or implement any specific interoperability technology. Instead the data-providers can ideally continue using their tools and publishing their resources according to their own standards, as long as common and internationally recognized formats are used (EUROGEOSS). In this way, the brokering approach

loosens the necessity for implementing a common data model and exchange protocol, by providing the necessary mediation and transformation functionalities (Nativi et al. 2015).
In recent years, GEOSS, in particular the EuroGEOSS project (http://www.eurogeoss-broker.eu/), has developed a brokering framework with the purpose of making heterogeneous resources from a variety of published Data Providers commonly discoverable and accessible to the user community (Nativi et al. 2013). The brokering approach is now implemented into the GEOSS common infrastructure (GCI) called GEO Discovery and Access Broker (DAB), and builds on a broker for each main functionality: discovery, access and semantic interoperability (Nativi et al. 2015).
Mature data integrators within AtlantOS, like SeaDataNet, are already making use of the GEOSS brokering services. For example, the XML encoding output of the SeaDataNet common data index service has been upgraded to the new SeaDataNet ISO 19139 Schema in order to comply with the EU INSPIRE Directive Implementing Rules. This conversion was amended using the GEO-DAB brokerage service that also plays an important function for the GEOSS portal (Seadatanet-Newsletter 2015)

The AtlantOS WP7 is closely following the evolution of the GCI and interacts with GEOSS in order to explore how AtlantOS as a community can contribute and make best use of GEOSS services for technical solutions to improved harmonization as well as for dissemination of AtlantOS data. (A report of GEOSS initiatives is planned for D7.8 (mo36)

Data harmonization through Data policies

All oceanic monitoring programs produce information and research products, which can provide the scientific basis for decision-making aimed at understanding and responding to challenges of the Atlantic Ocean such as climate change, pollution and natural hazards. It is therefore reasonable to infer that the projects and programs, which are funded to collect these data, bear a particular responsibility to both the global scientific community and to the society within, which they operate. Thus, open access to data collected under the AtlantOS framework is a prerequisite, which is ratified in the Data Management Plan (DMP) for AtlantOS (AtlantOS_DMP 2015).
Within AtlantOS there is a clear consensus to encounter the consequently imposed responsibility to comply with legal and ethical frameworks and ensure open access to the data collected as part of AtlantOS. Therefore, additionally to the important technological standardization steps, AtlantOS has undertaken to set up a common DMP. This DMP complies with the Open Access policy of H2020 and the rules of the open research data pilot. The DMP sets the framework for data handling and dissemination of data procured with AtlantOS funds.

The DMP sets the framework for harmonization of data produced during the AtlantOS project and includes essential guidelines for 1) Standardization, 2) Data exploitation and reuse strategy 3) Principles of access and sharing. The first DMP was installed in September 2015 (AtlantOS_DMP 2015).

The latest version of the DMP is available here: https://www.atlantos-h2020.eu/download/deliverables/11.2%20Data%20Management%20Plan.pdf.

The DMP is an evolving document and will be changing as the work of the AtlantOS, in particular WP7, progresses. An updated version the DMP will be available for the next reporting periods implementing the current outcome of WP7, which is mentioned previously in this report. An analysis of the current version of the AtlantOS DMP can be found as deliverable D7.3 https://www.atlantos-h2020.eu/structure/deliverables/

Data harmonization for the whole Atlantic Ocean

The AtlantOS **data integrators** and **data networks** represent mature and advanced data handling entities with long-term experience and a high degree of individual standardization. In general, many of these networks and integrators represent the state-of-the art of data management in oceanic data in Europe. However, Europe is only one of the four continents that border the Atlantic Ocean. Although the AtlantOS networks and data integrators deals with data resources from the entire Atlantic Ocean, there are resources bound in North- and South-America and Africa, which AtlanOS should also be compatible with. AtlantOS is in general working on strengthening the collaboration with the other countries around the Atlantic and in this regard collaboration with NOAA (USA) and MEOPAR (Canada) was initiated. The present barriers for trans-continental sharing of data include lack of organizational resources, insufficient institutional policies and support, disparate formats and file types, insufficient attention to the scientific imperative to share data, large amounts of legacy data with undocumented lineage, and a need for more user-friendly, certified long term data repositories willing to archive and disseminate Atlantic research data. In the light of these issues a workshop organized by AtlantOS, getting WP 7 and WP 8 involved, is currently planned for the first quarter of 2017. First contacts have been established to data managers in the US (NOAA) and South Africa (University of Cape Town) and topics that shall be addressed are long-term archiving and quality control and assurance (QA/QC) of real-time data. The later issue has been under discussion for over ten years on both sides of the Atlantic. With the NOAA supported initiative QARTOD a long-term process had been established that fully supports the GOOS Framework on Ocean Observing. During the last years, QARTOD is operating under the auspice of US IOOS that coordinates federally-owned observing and modeling systems and develops and integrates non-federal observing and modeling capacity into the system in partnership with IOOS regions. The expected outcome of the workshop will be to set-up robust links across the Atlantic to harmonize the principles of QA/QC for current, mature variables and work on the further development of procedures for other variables. This will also touch upon the discussion of Essential Ocean Variables (EOVs) and Essential Biodiversity Variables. With that the link to GEOSS in particular the new MBON (Marine Biodiversity Observing system) initiative of GEOBON can be established and the related activities supported.

In this way, AtlantOS shall continue to improve the foundation for sharing of Atlantic Ocean data in a fully international context.

References

AtlantOS_DMP (2015) AtlantOS_Data_Management Plan D11.1 https://www.atlantos-h2020.eu/download/deliverables/11.2%20Data%20Management%20Plan.pdf. Accessed 01.09.2015.

EUROGEOSS http://www.eurogeoss.eu/broker/Pages/AbouttheEuroGEOSSBroker.aspx. Accessed 01.09.2016.

Nativi S, Craglia M, Pearlman J (2013) Earth Science Infrastructures Interoperability: The Brokering Approach. IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING 6

Nativi S, Mazzetti P, Santoro M, Papeschi F, Craglia M, Ochiai O (2015) Big Data challenges in building the Global Earth Observation System of Systems. Environmental Modelling & Software 68:1-26

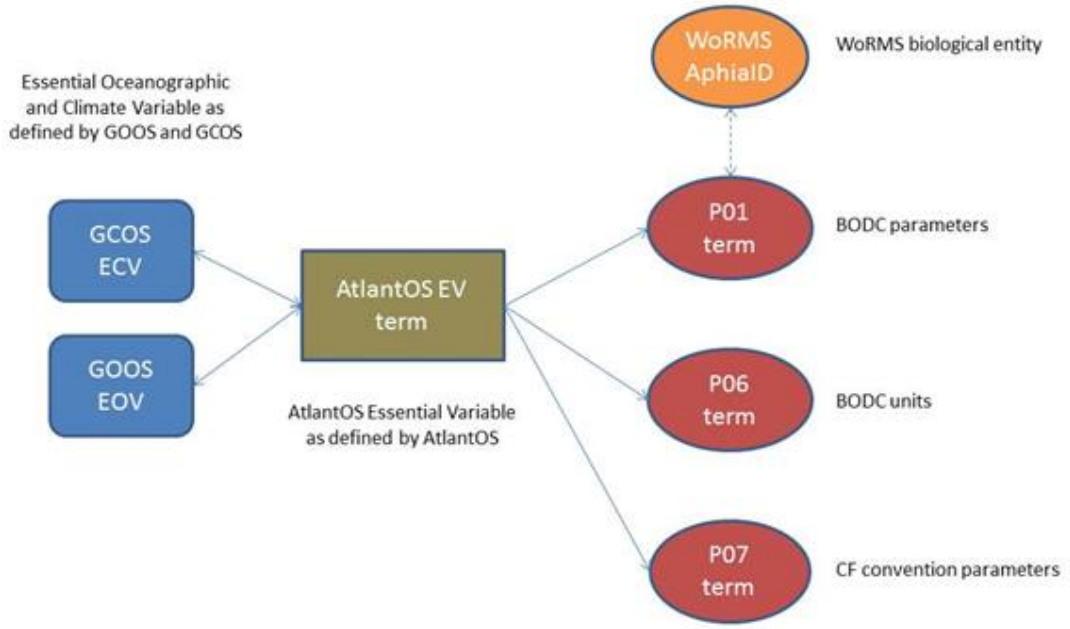
Seadatanet-Newsletter (2015) http://seadatanet.maris2.nl/newsletter.asp. Accessed 01.09.2015.

Appendices

Appendix 1: AtlantOS Data landscape and EOVs synthesis

			WP2 - ship based observing networks					WP3 - autonomous observing networks					WP4 - coastal observing systems				
			GOSHIP	SOOP	CPR	Fish + plankton survey	Seafloor mapping	Argo	Glider	Drifter	Oceansites	EATN	Tide Gauges	Ferrybox	FOS (RECOPESCA)	Coastal profilers	Fixed moorings
EOV	Physics	Temperature	M	A	A	A	A	M	100%	100%	M	A	A	M	M	M	M
		Salinity	M	A	A	A	A	M	100%	4%	M			M	M	M	M
		Current	M					A	100%	100%	A						A
		Sea Level											M				
		bottom depth					M	A				A					
		Air temperature									M		A	A			A
		Air humidity									M			A			A
		Atmospheric pressure								100%	M		A	A			A
		Wind speed									M		A	A			A
		Wind direction											A	A			A
		Rainfall									A						A
		Waves									A						
		Radiative fluxes									A			A			
	Biogeochemistry	Oxygen	M			A		M	80%		M	A		M		A	M
		Chla/Fluo	M		A			M	50%		M			M		A	M
		Nutrients (nitrate NO3,...	M					M	2%		M			A			
		Carbonate system (inorganic carbon)	M	M				not yet			A			M			
		Dissolved Organic Matt	A								A			A			
		Transient Tracers	M						25%								
		Nitrous Oxide	A														
		Turbidity							50%					M	A	A	M
	Biology/ ecosyste	Zooplankton			M	M											
		Phytoplankton			M												
		Species			M	M						M					
		Eggs and larvae			M	M											

Appendix 2: AtlantOS parameter harmonization matrix



The observed properties measured by networks will be labelled using terms from established vocabularies that are already being used by some of the AtlantOS networks (e.g. ARGO).
Observed properties will be labelled using the unique codes or preferred names from either the BODC **Parameter Usage Vocabulary (P01)** or **the CF (Climate and Forecast) Standard Names (P07)**. Units will be described using BODC Data Storage Units (P06).
It is recommended that observed properties from biological entities are labelled using the unique codes from **P01** and the internationally assured **AphiaID** (http://www.marinespecies.org/) from the WOrld Register of Marine Species (**WoRMS**).

An example of labelling the observed property and biological entity using P01 and AphiaID unique codes

Collection	Codes	Human-readable format
P01	SDBIOL06	*Abundance category of biological entity specified elsewhere*
AphiaID	115104	*Emiliania huxleyi (Lohmann) W.W.Hay & H.P.Mohler, 1967*

This is to facilitate their delivery into EMODnet Biology and will enable networks to label their data with the plethora of biological entities that could potentially be recorded in the Atlantic Ocean. We also recommend that networks use the WoRMS webservice (http://www.marinespecies.org/aphia.php?p=webservice) to adapt their own applications with standard WoRMS taxonomy. This will help assure that the biological entity identified is correct given the dynamic nature of classifying organisms.

To facilitate their discovery and to make them accessible to the observing networks, the vocabulary terms for AtlantOS key variables are published on the Natural Environment Research Council (NERC) vocabulary server (NVS2.0) under a **new vocabulary (A05)** which is searchable via the BODC vocabulary search user interface (https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search/A05/).

A05 vocabulary aggregates the essential climate and ocean variables acquired by the observing networks and helps overcome some of the complications related to using EOVs and ECVs, such as overlapping climate and ocean variables and the maturity of proposed EOVs (some of which are still immature). It also gives the networks the flexibility to request user-specific terms for observed properties in addition to the generic terms provided. The AtlantOS key variables in A05 will be mapped to their corresponding EOVs and ECVs once these lists are approved.

Summary of standard vocabularies and identifiers used in the AtlantOS parameter matrix

Collection	Title	Governance	URL
P01	BODC Parameter Usage Vocabulary	British Oceanographic Data Centre	http://vocab.nerc.ac.uk/collection/P01/current/
P07	Climate and Forecast Standard Names	Climate and Forecast Standard Names Committee	http://vocab.nerc.ac.uk/collection/P07/current/
P06	BODC data storage units	British Oceanographic Data Centre	http://vocab.nerc.ac.uk/collection/P06/current/
AphiaID	WOrld Register of Marine Species	WOrld Register of Marine Species	http://www.marinespecies.org/
A05	AtlantOS Essential Variables (EV)	British Oceanographic Data Centre	http://vocab.nerc.ac.uk/collection/A05/current/
TBC	GCOS Essential Climate Variables (ECV)	TBC	Not published yet
TBC	GOOS Essential Oceanographic Variables (EOV)	TBC	Not published yet